Giving Knowledge Bases a Voice – Towards Natural Language Generation from Structured Knowledge

Sandro Bauer
Computer Laboratory, University of Cambridge
sandro.bauer@cl.cam.ac.uk

Motivation

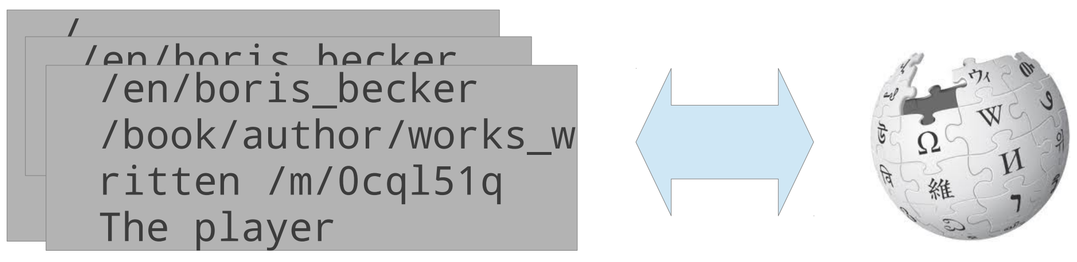
What does the Knowledge Base know about **Boris Becker**?

Hundreds of isolated facts stored as RDF triples

```
/en/boris_becker
/people/person/weight_kg 85
 /en/boris_becker /type/object/type
 /tv/tv_personality
  /en/boris_becker
  /book/author/works_written /m/0cql51q
  The player
```

But is this what end users really want?

Probably not. Ordinary users not familiar with the technical details of knowledge bases expect well-structured **natural-language output** in a coherent document – just as if they asked a domain expert!
This means we have to think about the way from structured knowledge in a KB back to natural language output.



Challenges

sentence level: mapping of abstract bits of meaning onto surface text (word choice, word order, grammaticality)
paragraph level: coherence, length, style, understandability, information order
document level: document structure, choosing appropriate sub-graphs to generate text from

Research ideas

- Explore what bits in the graph to use for generating text
- Learn generation models based on user feedback
- Learn a model for **joint** NLG and KB population

Research Vision

Can we use the abstract knowledge contained in a KB to create nicely structured human-readable summaries?

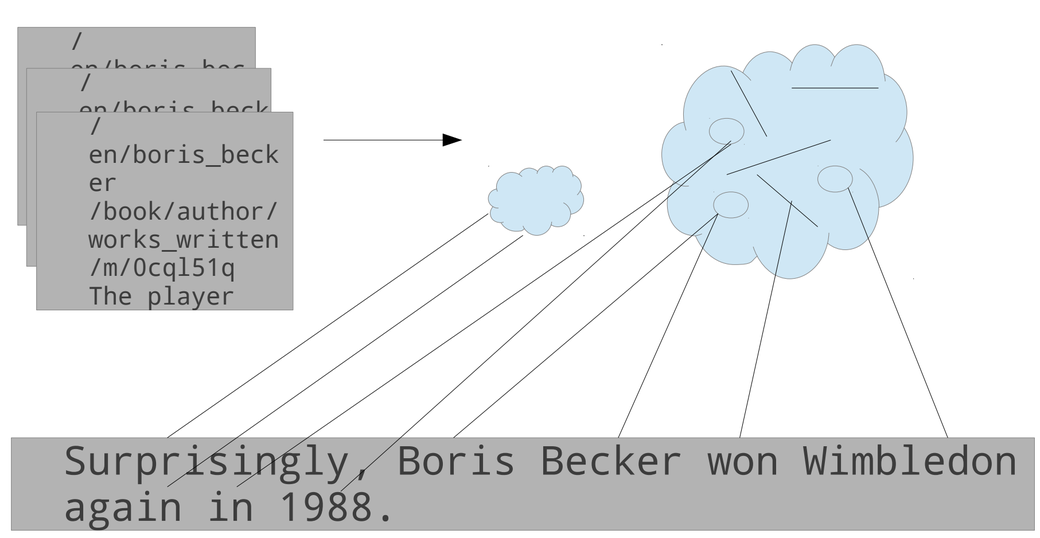
- Output in multiple languages
- Output tailored to the target audience
- Long vs. short versions of the same input
- More or less background information

Can we even go the other way round and inject new knowledge into the KB?

All-in-all: Make a KB a component as transparent as ever possible to the end user

Current state

- Research is in its very early stages (only started in late April 2012)
- Ground the project in Freebase and restrict ourselves to a simple model initially, making bag-of-words and bag-of-concepts assumptions
- Extend an LDA-style topic model: words are now generated from semantic bits in the graph
- Later include language models or other suitable components



- Unclear what bits of the KB to use for generating text
- Unclear what search terms to use for retrieving training data from the Web